Research & Methods
ISSN 1234-9224 Vol. 19 (1, 2010): 51–92
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw
www.ifispan.waw.pl
e-mail: publish@ifispan.waw.pl

Processing, Archiving and Dissemination of ESS data. The Work of the Norwegian Social Science Data Services

Kirstine Kolsrud, Hege Midtsæter, Hilde Orten, Knut Kalgraff Skjåk and Ole-Petter Øvrebø
Norwegian Social Science Data Services

Ever since the advent of the social science data archive movement, the work to reduce financial, technical, legal and administrative barriers between users and data resources has been a recurring theme, indeed a driving force behind the development of many of these institutions. Deploying a *data life cycle* approach, this article is an attempt to illustrate how the endeavour to reduce these barriers is tackled in the context of the biannual European Social Survey (ESS). The authors argue that the ESS' organisational structure, in which the Central Coordinating Team constitutes the backbone, is an important factor in overcoming barriers previously caused by lack of standardisation and harmonisation in cross-national surveys of this kind. Moreover, the introduction of cutting-edge data access arrangements, has lowered the legal and institutional barriers between ESS data producers and users substantially. This includes giving users without access to more sophisticated statistical packages the opportunity to browse and analyse data through the online data distribution tool Nesstar. However, the authors also suggest that the cumulative nature of surveys such as the ESS, poses a counterforce to the process of reducing barriers, challenging the data archives to seek new ways of improving the structure and design of their main dissemination channels. The authors are all involved in the ESS at the Norwegian Social Science Data Services (NSD). NSD is one of the seven scientific partners in the project and has served as data archive for the ESS since its inception.

Key words: European Social Survey (ESS), archive movement, barriers to data, data life cycle, data preparation, data archiving, data processing, data and metadata dissemination.

1. INTRODUCTION

1.1 NSD the official archive of the ESS

The Norwegian Social Science Data Services (NSD), has been the official data archive for the European Social Survey (ESS) since the first wave of the survey in 2002, and is one of seven scientific partners in the central coordination of the ESS. NSD is a national multi-disciplinary research service facility and one of the largest archives for research data in the world. In addition to the provision of research data, NSD offers a variety of services to researchers in Norway as well as internationally. NSD serves as a resource centre, assisting researchers with data gathering, questionnaire design, social science data analysis, methodology, privacy issues and research ethics.

The primary objective of NSD is to improve possibilities and working conditions for empirical research that is primarily dependent on the access to data. To fulfil this objective, NSD works to reduce financial, technical, legal and administrative barriers between users and data resources.

1.2 The archive movement

The history of social science data archives is not all that long, and the growth of archives has been compressed into a period of little more than a decade (Miller 1976). The social science data archiving movement began in the 1960s within a number of key social science departments in the United States which stored original coded interview data deriving from academic surveys (UKDA Home page). The establishment of archives for raw data from polls and surveys formed the basis for institutions like The Roper Center and the Inter-University Consortium for Political and Social Research (ICPSR). The movement spread across Europe and the Zentralarchiv in Germany (today GESIS Data Archive for the Social Sciences), the Steinmetz Archive (now part of DANS) in the Netherlands are examples of early European establishments (Rokkan 1976).

Later in the 1960s and early 1970s the UK Data Archive (UKDA), NSD and other archives were set up. Overarching organisations to enhance collaboration between the archives like Council of European Social Science Data Archives (CESSDA) in Europe and the International Federation of Data Organisations for the Social Science (IFDO) world wide did also emerge in the early 1970s.

The evolution of social science data archives can be ascribed to a combination of factors. Stein Rokkan, one of the key architects and initiators of NSD, saw the data service of social science archives as a response to the challenge of two parallel developments; one intellectual and one technological (Rokkan 1976). On the intellectual side, the emerging focus on empirical research and the increased prestige in quantitative methods created a great demand for data across large

populations. This spread rapidly from demography and economics to the entire spectrum of social sciences (ibid). Technologically, the computer revolution created a gap between the production and the distribution of data. As the statistical bureaus transferred their data sets, census sheets, register protocols etc. to machine readable media, the former transfer of data in the form of statistical tables, analysis and data sheets were no longer accessible from libraries and documents archives as these were still applying man-readable media. Hence the agencies responsible for distribution to the community of social scientists had no means of meeting the new demands for data from the user community as they could only handle man-readable data (ibid).

This mis-match between the increased demand for mass data from social scientists and the emerging methodological revolution in the social sciences and the lack of availability of information collected by government agencies, made room for the development of new infrastructure agencies. Furthermore, as the methodologies offered alternatives to the classical total enumerations of official data collections, the demand for secondary analysis from market research and standardised sample surveys also increased. In many instances the market research firm or the survey agency proved more flexible than the governmental agencies and were more capable of producing large quantities of data for the social sciences. The emerging data archives did also reflect this situation as some of the first organisationally distinct archives were established on the basis of raw data from polls and surveys. These first archives based on raw data from polls and surveys set off further efforts to build broader range agencies for the transmission of data for research on data from both governmental agencies and survey agencies (ibid).

In Norway, the Norwegian Social Science Data Service (NSD) was established in 1971. Financed by the Norwegian Research Council, NSD were, in contrast to other early established archives, based on the development of databases. Within the programme of electoral studies that was launched already in the 1950s, efforts to link up data from election statistics, party membership records, censuses and tax returns and a wide variety of official statistics were initiated. This was the start of what later developed into the comprehensive Commune Data Bank that was to become one of the cornerstones of the NSD. Another distinctive feature of NSD as an archive was, and still is, the ambition to be more than a distributor of machine readable data, and also to offer services of other computer based tools, teaching packages and workbooks for computer analysis as well as training in software systems (Rokkan and Henrichsen 1976).

1.3 The role and the work of the data archives

The overall objective of the data archives has been to make empirical data available for analysis. This means minimising any kind of barrier the end users may meet in their quest for empirical data. It includes preservation of data physically and to ensure the migration of data to timely data storage and formats. Depending on the nature and sensitivity of the data, access has traditionally been either at the physical location of the data or by distribution on magnetic tapes, discs, CD-ROMs etc.

Although the reservoir of data to be safely sorted for future use and re-use still is the foundation of the data archive, new technologies have in many ways placed new demands and also new opportunities for the archives.

For example, the storage capacity and variety of storage mediums have changed considerably. Instead of asking for subsets of data on magnetic discs, large datasets, even population data, can now be made available over the Internet

Furthermore, communication technology in general and the Internet in particular have changed the way end users and data archives communicate. Users can, in most cases, browse data catalogues on the Internet, search for particular datasets of interest and either order by e-mail or download the data directly from archive websites.

In addition, the combination of powerful PCs being available for the individual end users, combined with the development of user-friendly advanced statistical software, enables the users to perform increasingly complex analysis on larger and more complex data sets (Kolsrud 2007).

These developments naturally affect the role, scope and the day-to-day work of the data archives.

The Internet has for example gained a key role in the dissemination of survey data and most of the major cross-national surveys like the European Social Survey (ESS), International Social Survey Programme (ISSP), the Comparative Study of Electoral Systems (CSES), the European Value Study (EVS) and the World Value Survey (WVS) all make their data available over the Internet. In fact, end users do increasingly expect to have data accessible within a few mouse-clicks (Kolsrud 2009).

However, the use of the Internet to access data does also place more responsibility on the individual researcher to seek out the relevant and necessary documentation, securing correct use. The importance of well structured and documented data and metadata is thus more imperative than ever and is a core responsibility of the data archives as distributors of data.

As will be presented in the following chapters, the traditional distribution of survey data centring around two pillars; the data file and it's codebook, is also changing and we note that the traditional distribution of data and codebooks are replaced by new forms of dissemination, where data, metadata and analytical tools

are integrated. All in all, the increased use of the Internet places a clear demand on data archives to develop user-friendly and comprehensive websites equipping the research community with as complete and self explanatory sites as possible.

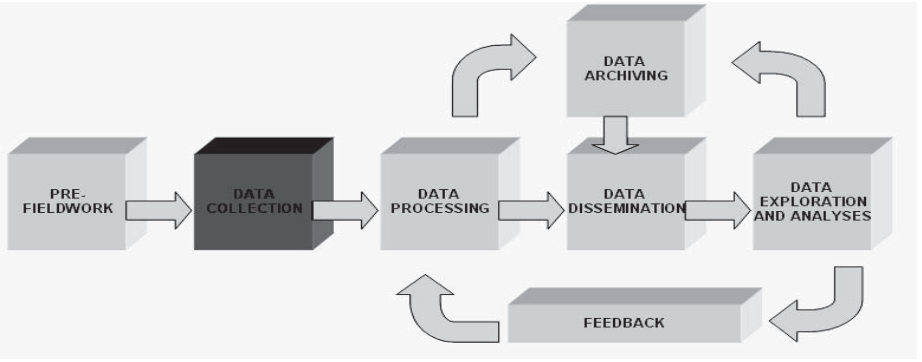
The role of the data archives as gateways to research data is rather well established and recognised in the research community. However, we would still like to argue that the internal work of the archives is not that well known. What takes place at the archive from the preparation and reception of data, to the disseminated and documented data files is still a bit of a "black box" for the end users.

In the following chapters we will address NSDs work with the ESS in the context of reducing barriers to data and in doing so shed some light on the "black box". We will present NSD's work with the ESS throughout the different stages in the data life cycle, from data preparation, ingest (data reception), processing, archiving and dissemination.

1.4 The Data Life Cycle

The data life cycle may be presented visually like the figure below:

Figure 1.1 The Data Life Cycle



As one of the seven scientific partners of the ESS, NSD has been in the unique position as an archive to be present in almost all stages of the project, from the pre-fieldwork period to the feedback from the users after they have explored and analysed the data.

The first stage, the Pre-Fieldwork stage, is where NSD's role is to provide the National Teams with instructions and specifications on data and metadata delivery. This is an important part of harmonising and standardising the data. All National Teams have access to and are required to apply the same specifications as provided from the ESS Archive Intranet, which we will get back to later in the article.

Stage two in the life cycle, is the Data Collection. This stage is completed by the National Teams and survey organisations, following the guidance of the Central Coordinating Team of the ESS. The data archive does not have an active role in this stage.

The next stage, the Data Processing, is where some of the core tasks of the data archive lies. It is in the data processing stage of the data life cycle, NSD harmonises the data sets which are deposited from the different National Teams. The data is checked for logical errors, filter errors and for breaches of the ESS standards and specifications. At this stage, communication with the National Teams is vital to the production of the final datasets. An important part of this stage is also the checking and processing of the documentation accompanying the data.

The fourth stage, is Data Dissemination. Here, the data archive distributes the datasets and documentation. In the case of the ESS, both the country-specific and integrated data files are made available from the ESS Data website (http://ess.nsd.uib.no). Together with the data files, all documentation and metadata is distributed. Once the data has been made available, the users are free to use the data for analyses. There is also an important feedback-loop in the life cycle. The exploration and analyses done by users, may reveal errors in the data. The feedback on these errors, leads to new and improved versions of the data and metadata. This is an important part of the validation and quality enhancement of the ESS.

The Data Archiving stage, including long-term preservation and timely storage, is a continuous and key task for the Archive.

In the following chapters we will argue that the overall aim of reducing barriers to data in many respect is inherent in the ESS as a project, and that NSD, in its role as scientific partner in the survey, has had a unique opportunity to follow the full data life cycle and hence implement enhanced procedures and specifications into the archiving and dissemination work.

2. BREAKING DOWN BARRIERS BETWEEN USERS AND DATA PRODUCERS, STANDARDISING THE PRODUCTION OF DATA

2.1 Barriers to data

One of the main objectives of the ESS has been to reduce the barriers between users and producers of data. In the earlier days of quantitative social science little was done to overcome the many difficulties of easy access to collected data, i.e. linguistic, cultural and institutional barriers. As social science research has matured, the principle of allowing others to replicate methods being used in different projects has become subject to more encouragement and facilitation. This has to a large extent been aided by the introduction of cutting edge data access arrangements.

The ESS has from the start been defined as a public good and aimed to achieve transparency and easy accessibility. The whole research community gets access to the data simultaneously, and so neither the principal investigators nor the question module design teams in the ESS have any guarantee of being able to be the first to publish their findings. As we will get back to later, ownership rights is one of the barriers to free and immediate access to data. With the simultaneous distribution, ownership of the ESS is effectively distributed between all users.

Kolsrud, Skjåk and Henrichsen (2007) identify what is commonly considered to be the most common barriers to free and comprehensive access to data. They divide them into three main types, as can be seen from table 2.1: Legal and institutional, organisation of data and documentation, and cultural.

Table 2.1 Barriers to free and comprehensive access to data

Legal and Institutional	Privacy/confidentiality; ownership rights/embargoes; pricing systems
Organisation of data and documentation	Lack of standardisation of variables; poor quality and lack of standardisation of metadata; lack of transparency; out-of-date information systems
Cultural	Customs/habits; attitudes, languages

The primary obstacle to free access to individual-level micro data is said to be the legal barriers as a result of privacy concerns. When individuals can too easily be identified in the data, restrictions on access are needed. However, for most general population surveys this does not apply, as the data is more often than not anonymised. Thus, when dealing with anonymised datasets, strict data dissemination rules are hard to justify with confidentiality being the only reason (Dale and Trivellato, 2002).

Another major barrier is the fact that some principal investigators often claim the right to impose strict constraints on the use of 'their' data by others. This may even be the case when the funds for the data collection are coming from the public purse, something that should in theory define the data as a public good. Data is also often only accessible when paying a considerable fee, which is sometimes the case with accessing data from statistical offices.

The idea of intellectual ownership of data is often an institutionally – or culturally – defined concept. According to Kolsrud, Skjåk and Henrichsen "If large scale, publicly-funded research projects are to justify their funding, they must do so on the basis of free and equal access to the data for all". The ESS has always been considered, and treated as, a public good. This means that the data are made available

to all simultaneously, without regards to whether they are a part of the ESS team or not. This principle allows the casual user to access and use the data at the same time as the ESS Central Coordinating Team, the question module design teams and the National Coordinators, and thus no one has the advantage of being first.

Another reason making this free and unlimited access possible in the ESS, is the fact that all published data are anonymous. It is the National Coordinators' responsibility to make sure the data they deposit to the archive is in accordance with their country's data protection regulations and privacy laws.

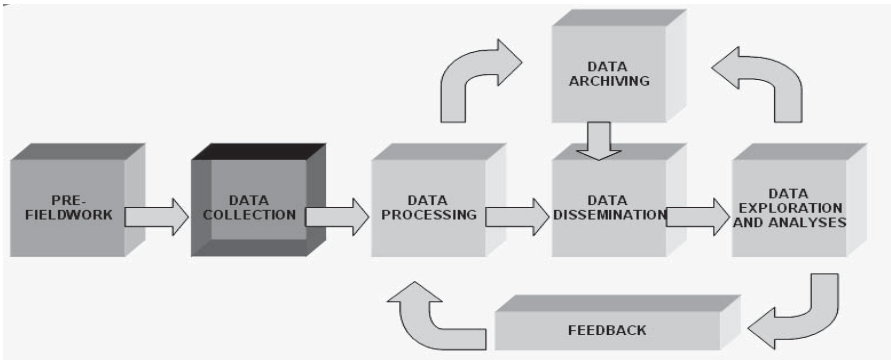
The collective nature of the ESS has by definition meant that most of the legal, institutional and cultural barriers have been eliminated, and so the main challenge has been to minimise the obstacles stemming from the organisation of data and documentation, which are all related to the production of data and documentation and how they are distributed to users. As table 2.1 shows, these barriers can appear in many forms, such as different variable names for the same variables, poor quality, lack of standardisation or even completely absent metadata, and the poor upkeep or lack of transparency of supporting information systems.

2.2 Standardising the production of data

One of the instruments for reducing barriers of an organisational nature is to produce standardised and harmonised data and metadata.

In the ESS, all participating countries are required to implement the survey according to the same set of specifications. The front end of the process, which involves a standardised survey design and data collection, has been given much thought. The same applies to how the subsequent data should be organised, standardised and made available to the end user. In the Pre-Fieldwork stage, the ESS Archive Intranet is the most important part of the archive's tools.

Figure 2.1 The Data Life Cycle, Pre-fieldwork stage.



The Archive Intranet is a "workbench" where specifications, standards and processing tools, as well as data files and documentation, are shared between the National Teams and the Central Coordinating Team (CCT), as well as within the CCT. The site contains all the information and documents needed in order to produce standardised and harmonised cross-national data files that eventually are published at the ESS Data website. It is here the National Teams download all the documents they need, like the Data Protocol, SPSS and SAS dictionaries and standards for post-coded variables.

Figure 2.2 The ESS Archive Intranet website

The ESS Archive Intranet is central in all stages of the production of data and metadata. The National Teams download the standardised specifications and subsequently deposit their data files and documentation to the site. When NSD have finalised the processing and quality control of the files, the National Teams download the harmonised national data files for validation. To enhance transparency, the National Teams have also access to the programmes that have been used by NSD during the processing.

The ESS Data Protocol is the key specification document available for the National Teams. The Data Protocol helps to achieve cross-national uniformity in data delivery, as it gives specifications for coding of data, the production of and the delivery of data files and other electronic deliverables. It contains sections

on the procedures for collaboration between the National Teams and the archive at NSD, electronic deliverables, principles of variable definitions, standards and classifications, and country-specific, identification and administrative variables. In addition to this, its largest part is a detailed coding plan which defines the variable names, answer categories, whether numeric or alphanumeric codes are used, and detailed routing instructions consistent with the instructions given in the source questionnaire. The routing instructions are supplemented by a flowchart.

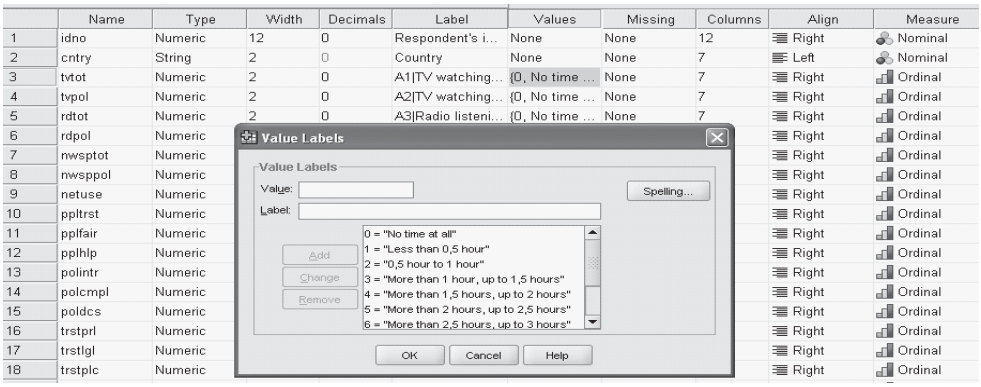
Figure 2.3 The ESS Data Protocol, coding plan

Table F.1a. Data file 1: Main questionnaire, section A. Idno and Cntry.

Qno	Name	Label	Format	Values	Categories	Comment
	IDNO	RESPONDENT'S IDENTIFICATION NUMBER	F9.0			See section E.2
	CNTRY	COUNTRY	A2			See sections E.1.1, E.2
A1	TVTOT	TV WATCHING, TOTAL TIME ON AVERAGE WEEKDAY	F2.0	00	No time at all	A1: Ask all Go to A3
				01	Less than 0,5 hour	
				02	0,5 hour to 1 hour	
				03	More than 1 hour, up to 1,5 hours	
				04	More than 1,5 hours, up to 2 hours	
				05	More than 2 hours, up to 2,5 hours	Ask A2
				06	More than 2,5 hours, up to 3 hours	
				07	More than 3 hours	
				77	Refusal	
				88	Don't know	
				99	No answer	
A2	TVPOL	TV WATCHING, NEWS/ POLITICS/CURRENT AFFAIRS ON AVERAGE WEEKDAY	F2.0	00	No time at all	Ask A2 if A1=01–07,77,88
				01	Less than 0,5 hour	
				02	0,5 hour to 1 hour	
				03	More than 1 hour, up to 1,5 hours	
				04	More than 1,5 hours, up to 2 hours	
				05	More than 2 hours, up to 2,5 hours	
				06	More than 2,5 hours, up to 3 hours	
				07	More than 3 hours	
				66	Not applicable	
				77	Refusal	
				88	Don't know	
				99	No answer	

The ESS variable names, labels and codes incorporated within the coding plan are also copied into programmes and "empty" data files (dictionaries) in SAS and SPSS. Information on these programmes are also to be found in the Data Protocol. For countries using CAPI, these can be used in building the CAPI programmes. They are also used by the National Teams to build data entry programmes and the national data files.

Figure 2.4 SPSS data dictionary



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	idno	Numeric	12	0	Respondent's i...	None	None	12	Right	Nominal
2	cntry	String	2	0	Country	None	None	7	Left	Nominal
3	tvtot	Numeric	2	0	A1|TV watching...	{0, No time ...	None	7	Right	Ordinal
4	tvpol	Numeric	2	0	A2|TV watching...	{0, No time ...	None	7	Right	Ordinal
5	rdtot	Numeric	2	0	A3|Radio listeni...	{0, No time ...	None	7	Right	Ordinal
6	rdpol	Numeric							Right	Ordinal
7	nwsptot	Numeric							Right	Ordinal
8	nwsppol	Numeric							Right	Ordinal
9	netuse	Numeric							Right	Ordinal
10	ppltrst	Numeric							Right	Ordinal
11	pplfair	Numeric							Right	Ordinal
12	pplhlp	Numeric							Right	Ordinal
13	polintr	Numeric							Right	Ordinal
14	polcmpl	Numeric							Right	Ordinal
15	poldcs	Numeric							Right	Ordinal
16	trstprl	Numeric							Right	Ordinal
17	trstlgl	Numeric							Right	Ordinal
18	trstplc	Numeric							Right	Ordinal

The ESS uses acknowledged international standards for the coding of verbatim recorded variables, such as occupation, industry, country codes (citizenship, country of birth, etc.) and language. Information on the use of these standards is found in the Data Protocol, while the coding standards themselves are available for download from the ESS Archive Intranet.

Figure 2.5 Archive intranet, international standards and coding frames

The standards used are compiled for each round of the ESS and are treated as the only valid standards for that round. In addition the ESS has developed its

own coding frame for religion. The frame covers the seven largest religions in the world. It is based on country-specific coding schemes, which are later bridged into the ESS standard.

2.3 Survey documentation

Metadata is a large part of producing high quality data (Mohler and Uher 2003). The ESS has placed great emphasis on providing comprehensive and structured documentation for the end users. The fact that ESS is a complex and cross-national project which provides immediate access to data on the Internet, makes the accessibility and comprehensiveness of the documentation even more crucial.

Figure 2.6 National Technical Summary

European Social Survey

National Technical Summary 2008

Please fill in this form, save, and include it in the deliverables you submit to NSD

A1 Key persons and institutions	
Name and address of field work organisation	**A1.1 Data collector, fieldwork organisation(s)** Name:
Name of the person who provided the data to the archive.	**A1.2 Depositor** National Coordinator ❑ Fieldwork Organisation ❑ Name:
A2 Funding	
Full name of the country's funding agency or agencies.	**A2.1 Funding agency (agencies)**
The grant number(s) connected to the funding agency or agencies.	**A2.2 Grant number(s)**
A3 The collection of data	
Field work period (DD/MM/YY).	**A3.1 Date of collection** From: To:
Mode of data collection, main questionnaire *If PAPI used, please specify how the data were transferred to an electronic format*	**A3.2 Mode of data collection, main questionnaire.** • Computer assisted personal interview, CAPI ❑ • Paper and pencil interview, PAPI ❑ ○ Data keyed from questionnaire ❑ ○ Data optically scanned from questionnaire ❑
The language or languages in which the survey was conducted.	**A3.3 Language(s)**
Experienced interviewers: have previously done one or more interviewing projects *Inexperienced interviewers: interviewers with no previous interviewing experience*	**A3.4 Fieldwork procedures** A3.4.1 Interviewer selection • Total number of interviewers: _____ • Number of experienced interviewers: _____ • Number of inexperienced interviewers: _____

Just as the ESS data needs to be harmonised and standardised, it is equally important that the documentation is structured and standardised to be easily compared across the participating countries. In the ESS, the documentation requirements are made known to the National Teams prior to fieldwork. A National Technical Summary form, made available from the Archive Intranet, is the main vehicle for the collection of metadata from the National Teams (figure 2.6). It covers information about fieldwork dates, response rates, fieldwork procedures, funding, how the data was collected and interviewers. In addition, the National Teams are asked to give some country-specific contextual information such as the nature of the education system, the party system and the demographic composition of the population.

Furthermore, the National Teams are asked to deliver all the documents used during fieldwork, like interviewer instructions, advance letters, fieldwork instructions, the translated questionnaires, the contact form, show cards etc.

All the information given in the NTS is stored in a relational database which generates the ESS Data Documentation Report. The report contains a general study description in addition to the country-specific reports with information from the National Technical Summaries the National Teams have delivered. The information from the NTS is also supplemented with further documentation from the Central Coordinating Team and the ESS Sampling team.

All in all, the ESS metadata consists of more than 200 different documents per round, from the source questionnaires to the country-specific documents to the Data Documentation Report. These are all made available together with the ESS dataset. Thus the users are equipped with ample opportunities to analyse the data correctly.

3. PROCESSING AND ARCHIVING OF ESS DATA AND METADATA

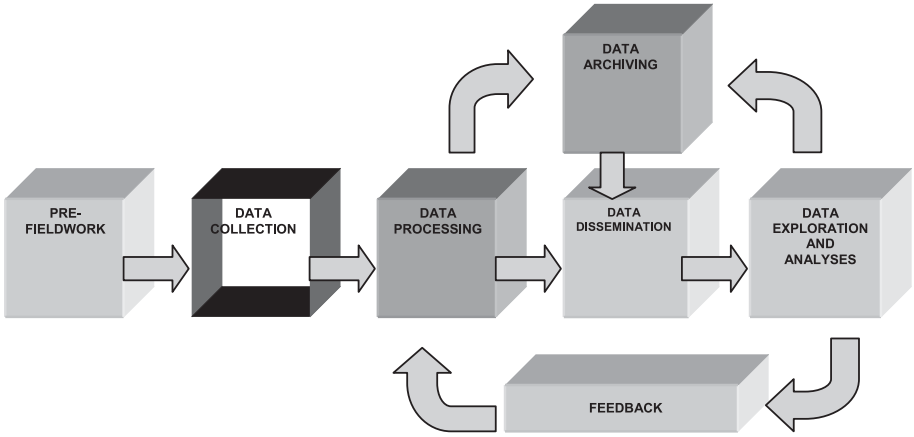
The next stages of the Data Life Cycle, Data processing and Data Archiving, represent the main tasks of the data archive. In addition to checks, corrections and harmonisation of data and metadata, these stages include archiving, preservation and maintenance based on feedback from users.

3.1 Principles for data processing

NSD's overall principle is to produce harmonised and standardised data files that balance two aims. Firstly, the data should be as user-friendly as possible. This means for example that the accuracy of data should be maximised, consistency of data should be established and the best data possible should be obtained. In addition, data should be as comparable across countries and time as possible. On the other hand, we believe that the data files should reflect the original reliability

and quality of the data. This means that the editing of data at the archive, as well as in each country, should be exercised with great caution (Skjåk 2007).

Figure 3.1 The Data Life Cycle, Data Processing and Data Archiving stages



One general example illustrates this dilemma from an end-user's point of view. If data are inconsistent and it turns out to be impossible to verify the true values of data, should the data be automatically cleaned according to the logic of the instruments, should they be edited according to what the data editors think is most likely, or should the inconsistencies in the data be retained? In the first two cases, the end user is spared irritating anomalies in the data. In the last, the data themselves document their quality, and the end user has the opportunity to consider how she/he should deal with the inconsistencies. As we will see later in this section, NSD has developed internal rules on how data should be edited in case of, for example, inconsistencies in routings.

Another important principle, is that data processing should be done in close collaboration with the National Coordinators and their teams. During the data processing stage each National Team has full access to the catalogue ("virtual workspace" containing all data and all programmes) where the processing of their data is actually taking place. All decisions about the editing of data are based on advice from the National Coordinators, who have first-hand knowledge of their national data, and the National Coordinators also have to approve the final drafts of the data files.

The last principle, is that the data processing should be consistent across data files and countries. In the processing of each ESS round, a team of five to six people at NSD are involved, and to ensure as coherent and consistent processing

as possible, a processing handbook specifying control and editing rules, and a set of common processing programmes have been developed. Decisions about editing of data that are not straight forward are discussed in the group. Consistency is also ensured by the balance between automatic and manual procedures. A common e-mail address is used by everyone to keep everybody updated on the progress. In addition to consistency, these internal procedures ensure internal transparency and response to National Teams, also when their primary contact person at NSD is out of office.

3.2 What is being processed

All in all, each ESS round consists of seven different data files and documents which are processed completely or partially by NSD:

1. Data from the core and rotating module questionnaire. This questionnaire includes two main sections, each consisting of 100-120 items; a 'core' module which remains relatively constant from round to round, plus two or more 'rotating' modules, repeated at intervals.
2. Data from the Interviewers' questionnaire, where interviewers for each completed interview fill in information about interview settings and respondent's cooperativeness.
3. Data from contact form (call records), where all contacts and/or contact attempts with all persons/addresses in the sample are registered and described.
4. Data from the questionnaire testing reliability of ESS instruments (MTMM analysis).
5. Sample data with information about sampling units, sampling probabilities etc. for all respondents in the sample.
6. Data files with parents' occupations, as text strings and/or ISCO codes.
7. The National Technical Summary Form (national documentation).

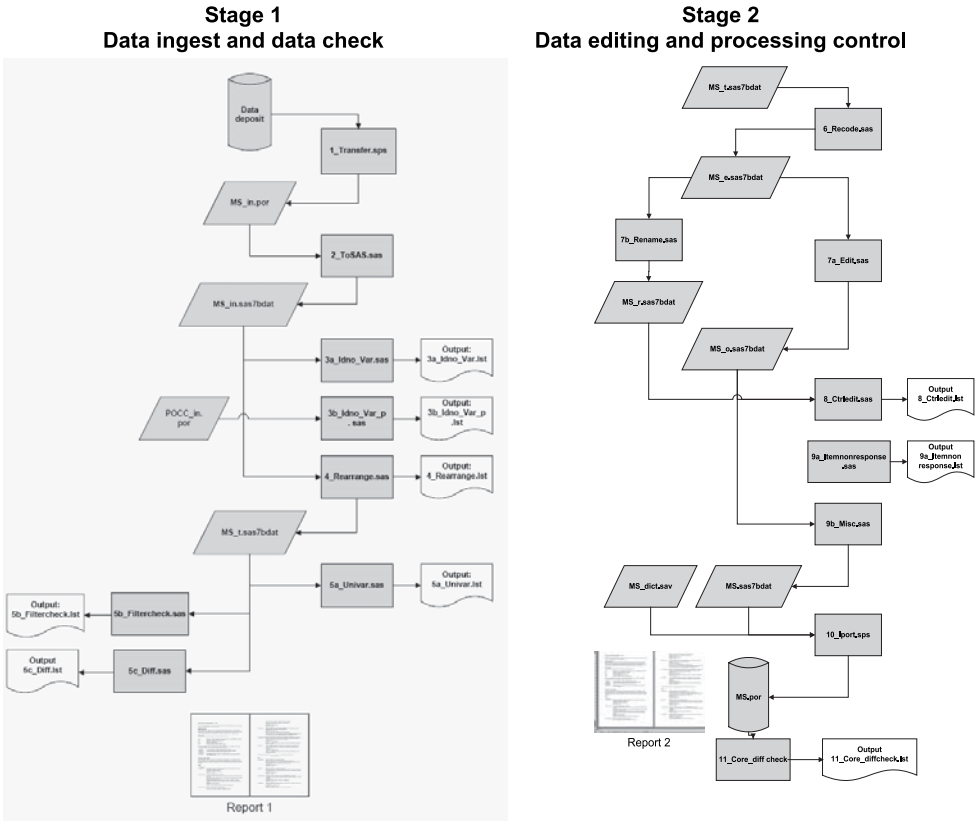
In the following, we will describe NSD's processing of the files, with emphasis on the data from the core and rotating modules, using ESS round 4 as use case.

3.3 Stages in the data processing

All in all, 13 data programmes are applied, from the files are deposited to the ESS Archive Intranet by the National Coordinators until the final draft files are ready for control and approval. Some of the programmes do automatic checks of the data files while other produce output to be controlled manually. The presentation of the processing stages will be illustrated by outputs from the different programmes. As mentioned in the previous section, all programmes, files and outputs are available for the National Coordinators, securing full transparency during the processing.

Each programme represents one distinct step in the processing, but we identify two main stages, each of them completed by a data processing report to the National Coordinators and their feedback (figure 3.2). The first stage consists of what we describe as *data ingest* and *data checks*, using programmes 1 to 5 in the processing flowchart. The second is described as *data edit, edit control* and *data approval*, using programmes 6 to 11.

Figure 3.2 ESS data processing flowchart



3.3.1 Stage 1 – data ingest and data checks

Data ingest involves the checking of formal attributes of the data files, while data checks controls for wild codes and inconsistencies in the data. The aim is to improve the data quality, the standardisation and the harmonisation as far as possible within the frames given in the processing principles mentioned above.

Starting with data ingest, the following is controlled for as soon as the data have been identified as readable and moved to the country's workspace:

Duplicate or missing identification numbers and consistency of identification numbers across files
If a data file has two data rows with the same identification number or data rows where the identification numbers are missing, this indicates that something is wrong with the data, for example that data from the same respondent have been entered into the data file twice. Users should be able to combine all data files from each round, using the unique identification number of each respondent as the key to the merging. If identification numbers are not consistent across all data files, such combination is impossible. It also indicates that some errors have been done either during fieldwork or in the initial production of data. The control of identification numbers is automatic. Figure 3.3A illustrates an example where two duplicates where spotted in a file.

Figure 3.3 Examples of output from data ingest programmes

A. Duplicate identification numbers

```
Norway: duplicate or missing IDNO in Sample Data file

       IDNO    ROW

        258    168
        545    343
```

B. Identification of country-specific variables and additional variables

```
DIFF=DE name not present in ESS file

          DE      ESS
name      var     var   Variable label

EDLVDE    252       .   Highest Level of education, DE
EDLVFDE   314       .   Father's highest level of education, DE
EDLVMDE   321       .   Mother's highest level of education, DE
EDLVPDE   292       .   Partner's highest level of education, DE
PRTCLDE    35       .   Which party feel closer to, Germany
PRTMBDE    38       .   Member of which party, Germany
PRTVDE1    24       .   Party voted for in last national election, Germany
PRTVDE2    25       .   Party voted for in last national election, Germany
REGIONDE  367       .   Region, Germany
REFDE      31       .   Additonal variable Germany:Signed a referendum last 12 month
SPLOW1DE  361       .   N1|Place of interview East/West Germany
PRTSOWDE  362       .   N2|Where did your parents live before 1990
SPLOW2DE  363       .   N3|Where did you live before 1990
SPLOW3DE  364       .   N4|Where did you live last in Germany before 1990
SPLOW4DE  365       .   N5a|When did you move to West Germany
SPLOW5DE  366       .   N5b|When did you move to East Germany
```

Content checks

The next step of the data ingest is to control which variables are present in the data files. The data files deposited to NSD contain three types of variables:

- Variables that are identical across all countries (common variables). These variables should have the same names, formats and categories.
- Country-specific variables included in the ESS specifications or source questionnaires (ESS country-specific variables), for example party membership and region. These variables have the same stem in the name, but the country code is added as suffix, for example REGIONDE. The variables have country-specific formats and categories.
- Country-specific variables not included in the ESS specifications or core questionnaires (additional variables). These variables are not included in the official ESS data files, but are made available "as they are" in country-specific files.

The variables in the national files are automatically compared to the common variables of the source ESS files (produced by PROC CONTENT in SAS). This control reports missing common variables, eventually misspelled variable names, deviating variable formats, additional variables and the ESS country-specific variables. Figure 3.3B presents an extract from the output for Germany. The output confirms that the required country-specific variables about education, political affiliation and region have been delivered. We also see that several additional variables have been deposited.

Variable names

Since the ESS is a partly repetitive survey, a crucial aspect of the data production is to facilitate for cumulative files with comparable and correct data and metadata over time. The ESS use mnemonic variable names in order to have unique and consistent identifiers for variables used in two rounds or more (for example HAPPY). In the case of country-specific variables and in particular variables where political parties are categories, the content change rapidly in many countries. The checking of categories in country-specific variables is therefore considered as a part of the data ingest, since changes in the categories will result in changes of the variable names of the data files. For example, the name of the first version of the variable "Party voted for in last general election, Poland" was PRTVTPL, the second PRTVTAPL, and the third so far PRTVTBPL.

The data ingest is concluded by a programme assigning corrected variable names, formats and other formal attributes. Errors in the identification numbers are reported to the National Coordinators in the first processing report. The data ingest is followed by a set of procedures for data checks, found in programmes 3, 4

Figure 3.4 Examples of output from data checks of codes

A. Invalid codes in occupation (ISCO-88) – automatic control

```
Wild codes in ISCOCO, respondent IDNO

IDNO    ISCOCO

   9     5136
  12     5221
  30     5135
  32     5134
  35     5221
..
..
..
2649     5221


Wild codes in ISCOCO, frequency

              F24-25a|Occupation, ISCO88 (com)

                                    Cumulative    Cumulative
ISCOCO    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
  1120           2        0.67             2         0.67
  1130           3        1.01             5         1.68
  2250           1        0.34             6         2.02
  2511           1        0.34             7         2.36
..
..
..
  8342           2        0.67           296        99.66
  9442           1        0.34           297       100.00
```

B. Detecting missing code – manual control

```
Frequencies, Core and rotating modules file


VALUE      .    0      1      2     3     4     (5)     6     7     8     9
..
..
PVOLWRK    .    .    745     96   346   361      .     .     1     .     1
TWCOL20    .    .    305    523   203   155     (.)   363     .     1     .
TWCOL70    .    .    951    196    28    11     (.)   363     .     1     .
SGR2070    .    .    121    811   166   435      .     .     1    15     1
..
```

and 5 in the work flowchart. Here, the data are controlled for invalid, out-of-range, extreme and missing values, consistency between variables, structural consistency in routed variables and consistency over time in some key socio-economic variables. Like data ingest, the data checks consist of a combination of automatic and manual controls.

Wild codes – invalid, out-of-range, extreme and missing values

Errors in the values of variables are mainly due to failures in the keying of data, but also other errors might occur, for example that categories are left out from the questionnaires or that wrong standards are used in the post-coding of verbatim recorded variables. The most straightforward procedure in checking values is to identify the invalid and out-of-range values. These checks are done in two turns. First, variables using large standardised coding frames like ISCO-88 (occupation), NACE (industry), ISO-639 (language) and ISO-3166 (country) are automatically checked against the valid codes of the standards, and invalid values are listed. As shown in figure 3.4A the invalid values are reported on individual level with identification numbers as well as in a frequency table. Secondly, values, distributions and descriptive statistics in other ESS common and country-specific variables are checked partly manually, in order to spot extreme values, missing values and suspicious contributions. This is the most cumbersome step of the data processing, but it is a significant part of the data checks. Figure 3.4B illustrates an example where a missing category in a national questionnaire was spotted by the manual check.

Data consistency

Consistency of answers in attitudinal and behavioural questions is not controlled for. It is up to the analysts of data to identify and evaluate the extent of non-attitudes and inconsistency in behaviour, so consistency checks in the data processing is limited to the technical quality of data attributes. One example of inconsistency checked for is whether the respondent began in paid work before she/he was born, or as in figure 3.5A where the household grid reports that respondents are younger than their children or older than their parents. This could of course be the case in step relations and where the parent is a widow(er), but it is worthwhile to check whether the parent is 25 years younger than the child (IDNO=876 in figure 3.5A) or if it is due to errors in data registration.

Structural consistency

Structural consistency in routed variables is checked for in the core part of the questionnaires, i.e. questions blocks that are repeated in each round of the ESS. This step is also partly automatic and partly manual in the sense that programmes

Figure 3.5 Examples of output from checks of data inconsistencies, structural inconsistencies and inconsistencies over time

A. Data inconsistency

```
Second person in household: child/step/adopted older than respondent or parent/step
younger than respondent

IDNO    RSHIPA2    YRBRN    YRBRN2
  77        3       1964      1987
  79        2       1991      1990
  90        2       1985      1962
 876        2       1990      1965
```

<Explanation:
RSHIPA2:
2 = “Son/daughter (inc. step, adopted, foster, child of partner)”
3 = “Parent, parent-in-law, partner’s parent, step parent”>

B. Structural inconsistency

```
Filter error b20a-b20c if b20a = 1

 N of
 ases    CLSPRTY    PRTCLcc   PRTDGCL

 897          1          66     1-4,7,8
  19          1          66           9
=====
 916
```

<Explanation:
CLSPRTY: 1 = “Yes, feeling close to a party”.
PRTCLcc: 66 = “Not applicable for answering which party you are feeling close to”.
PRTDGCL: 1-4,7,8 = Substantive answer to how close you feel to that party.>

C. Inconsistency over time

```
Percentage difference ess round 3 - ess round 4

EDULVL

value     0.0    1.0    2.0    3.0    4.0    5.0    6.0    7.0    8.0    9.0
ess3      1.1    0.0   17.9   35.2    9.9   34.6    1.0    0.1    0.2    0.0
ess4      0.3    1.7   13.0   40.8    7.4   35.2    1.3    0.3    0.0    0.2
diff      0.8   -1.7    5.0   -5.6    2.5   -0.5   -0.3   -0.2    0.2   -0.2
```

do the checking, while the control of programme outputs and identification of action points are manual. The main aim with this check is to uncover either large numbers of respondents being incorrectly routed or systematic routing errors due to structural errors in the questionnaire or CAPI system. The example in figure 3.5B is the first type, where a question hasn't been asked to a large number of eligible respondents.

Consistency over time in selected background variables
Achieving correct and consistent coding of variables based on national, country-specific instruments and verbatim recorded answers is a major challenge in creating harmonised data in cross-national social surveys. In general, controlling the data quality of this group of variables requires thorough knowledge of social structures in each country and must primarily be done by the National Teams, researchers and analysts when the data have been published. However, as an extra control, NSD compare the frequencies of such variables with previous results for countries that have participated in more than one round of the ESS. Inconsistency in data over time could indicate that the national instruments and/or coding have been changed. Whether this is the result of an intentional improvement or not, it needs to be documented and made available to the end users. In the ESS, such comparisons are carried out for religious denomination, education level (respondent, partner, father and mother), industry, occupation (respondent and partner), 1st and 2nd language spoken at home, citizenship and country of birth (respondent, father and mother). An additional check in this programme is to control the correlations between occupations and education levels for respondent and partner.

Stage 1 is concluded by the first data processing report to the National Coordinators (figure 3.6) and their feedback on the issues listed in the report. When the final decisions on how to correct and upgrade the data have been set, re-coding of data is done by NSD, or the National Team delivers corrected variables or a complete data file replacing the original. Stage 1 is therefore often an iterative process because new data have to be controlled to the same extent as the data originally delivered.

3.2 Stage 2 – data edit, edit control and data approval

Data edit – recoding of wild codes and corrections of identification numbers
When the data issues in stage 1 have been sorted out, wild codes are corrected where the true values can be positively identified by the National Coordinator. This is often the case with invalid values in verbatim recorded variables like for example occupation, where the National Team can go back to the textual descriptions. All unresolved wild codes are set to "No answer" (codes 9, 99, 999 etc.), and this

Figure 3.6 Extract from the first data processing report, with results from the data ingest and data checks stage. References to figures 2 – 4 are added in parentheses in left column)

Data Processing Stage 1

Programs 1 to 5c_Diff.sas have been processed in Stage 1, and the results are documented in the tables below. Further progress depends on feedback from you, marked with # and underscore.
Please respond to our query by including your response in the tables below and return the report to NSD by e-mail. Please note that Stage 2 of the data processing will depend on National Team's feedback from stage1.

Programs 3 and 4 cont. (p.1 to 25 in Annex)	
IDNO (figure 2A)	# Duplicate IDNOs in Sample Data file for IDNOs 258 and 545. See p. 1 in Annex. Delete duplicate rows from Sample Data file?
ISCOCO (figure 3A)	# Wild codes 1120, 1130, 2250, 2511, 2512, 2519, 2521, 2523, 2531, 2541, 2542, 2545, 2551-2555, 2560, 3341, 3349, 3418, 3491, 3493, 5134-5137, 5164, 5221-5224, 6210, 6411, 7125-7127, 7144, 7217, 7234, 7236, 7237, 7243, 7350, 7450, 8114, 8213, 8214, 8267, 8342, 9442 appear in a total of 297 cases. See pp. 7-17 in Annex. Has occupation been coded into the correct standard [ISCO-88 (com)]? If not, please deposit new ISCOCO variable.
RSHIPA2 (figure 3C)	# There appears to be a logical error in one of the cases, as son/daughter/step/adopted/foster (category 02) has been coded as older than respondent. See p. 25 in Annex. Please comment.
Program 5a (p.26 to 50 in Annex)	
TWCOL20 TWCOL70 (figure 3B)	# No respondents coded in category 5 ("No work with other people last month") on these variables. It appears as if category 5 for these variables has been left out of the national questionnaire altogether. See p. 35 in Annex. Please comment.
Household grid: Child/step older than respondent or parent/step younger than respondent	# Please see page x to xx in the annex. For IDNO = 77, 79, 90, 876, 995 and 1010 there are some inconsistencies child/step/adopted is reported to be older than respondent or parent/step younger than respondent (See p. 12 in the annex): Second person in household: child/step/adopted older than respondent or parent/step younger than respondent IDNO RSHIPA2 YRBRN YRBRN2 77 3 1964 1987 79 2 1991 1990 90 2 1985 1962 876 2 1990 1965
Program 5b (p.51 to 51 in Annex)	
CLSPRTY PRTCLcc PRTDGCL (figure 3D)	# Filter error B20a (CLSPRTY)-B20c (PRTDGCL) if CLSPRTY=1. A total of 916 cases have been coded 66 ("Not applicable") on PRTCLcc despite having been coded 1 on CLSPRTY. Moreover 897 cases have valid values on PRTDGCL despite having been coded 66 on PRTCLcc. See p. 51 in Annex. Please comment.
Program 5c (p.52 to 53 in Annex)	
EDULVL EDULVLP (figure 3E)	# When comparing Round 4 data with ESS3, it seems category 3 has become bigger, whereas category 2 has decreased correspondingly in Round 4. See p. 52 in Annex. Does this change reflect real change or could it be due to different coding/bridging procedures in the two rounds?

is done in nearly all cases where keying or registration error is the source of the wild codes. Problems around identification numbers mentioned above are also corrected for in this step.

Data edit – assignment of missing values
As mentioned in 3.1, NSD has developed a set of principles on how data should be automatically edited when, for example, inconsistencies in routings occur. In addition, rules have been established to automatically assign missing values to all variables in general. Again, consistency, standardisation and harmonisation of cross-national data files is considered vital for the usability, at the same time as the data should reflect the original reliability and quality of data. The basic principles for the editing of inconsistencies in routings and assignments of missing values are as follows:

a. The data files can have five types of missing values
6, 66 etc. = Not applicable
7, 77 etc. = Refusal
8, 88 etc. = Don't know
9, 99 etc. = No answer, i.e. missing data not elsewhere explained
. (sysmis) = Variable not relevant, not deposited, or for other reasons omitted from the data file by the archive.

b. Not applicable is used only when data unambiguously confirm it.

c. Unresolved wild codes are set to "No answer".

d. Inconsistencies between "filter" variables and "substantial" variables: data in substantial variables are not edited, while data in filter variables are set to "No answer".
Example: If a respondent according to the data answers "No" when asked if she/he voted in the last general election, but also answers which party she/he voted for, the party variable is kept unchanged, while the first is recoded from "No" to "No answer".

e. Inconsistencies between substantial variables are not edited (cf. section on "Data consistency" above).
Example: If a respondent, according to the data, answers that he/she doesn't watch television at all on an average weekday, but also answers that she/he watches the news or programmes about political and current affairs more than three hours on an average weekday, both variables are kept unchanged.

Control of data edit

The programme assigning missing values throughout the data file marks the last major, standardised step in the processing of the national data. Before national data are returned to the National Coordinator for control and approval, the data editing described above is in itself controlled. If a variable has a different distribution after the editing than before, this is reported in the output from the control programme. Figure 3.7 illustrates a simple and very common example, where it is documented how missing values have been assigned in a routed variable (Party vote) based on the filter variable (Did vote). The upper table documents the editing itself, the lower table how this has influenced the distribution.

Figure 3.7 Control of the data edit

Estonia: Edited VOTE or Party

Nof Cases	input VOTE	input PRTVTEE	output VOTE	output PRTVTEE
513	2	.	2	66
194	3	.	3	66
12	8	.	8	66
3	9	.	9	99
=====				
722				

VALUE	.	1	2	3	4	5	6	7	8	9	10	11	66	77	88	99
VOTE	.	940	513	194	.	.	.	.	12	3	.	.	.	.	.	.
PRTVTEE	**722**	133	237	294	46	81	69	6	1	2	1	2	.	14	53	**1**
edPRTVTEE	.	133	237	294	46	81	69	6	1	2	1	2	**719**	14	53	**4**

<Explanation: VOTE: 2="Did not vote", 3="Not eligible to vote", 8="Don't know", 9="No answer".
PRTVTEE: 1-11=Political parties, 66="Not applicable", 77="Refusal", 88="Don't know", 99="No answer".>

Approval of data and integration in cross-national data files

Stage 2 is concluded by the second data processing report to the National Coordinators (figure 3.8). This report is normally much shorter than the first, reporting some unresolved issues from stage 1 and miscellaneous issues like how one should deal with respondents with a very high share of missing values (which is controlled for at the end of the processing when all missing values have been assigned).

Alongside the second processing report, follows a draft national data file to be controlled and hopefully approved by the National Coordinator. Before the data file is transferred to the National Coordinator, she/he has to agree upon a confidentiality agreement, stating that the draft file isn't disseminated outside the

National Coordinator's team. This is important for ensuring the ESS principle of "equal access for all".

Figure 3.8 Extract from the second data processing report

Data Processing Stage 2

Programs 6-11 have been processed in Stage 2. The processing has included actions agreed upon after the report in Stage 1.
Stage 2 includes program 7a_Edit.sas, which edits missing values and wild codes in all variables. The edits are documented in the Annex of this second report. The output of program 9a, showing the proportion of item non response in total as well as section-wise, will also be included in the Annex.

Unresolved issues from stage 1		
IDNO	**NSD comment**	# Duplicate IDNOs in Sample Data file for IDNOs 258 and 545. Has this duplication been corrected?
	Response from NT	Yes. An updated Sample Data file has been deposited to the ESS Data Archive.
ISCOCO	**NSD comment**	# IDNO 2124 has been recoded from 7125 to 7124 as agreed upon in previous correspondence. Please confirm.
	Response from NT	Ok.
EDLVAcc	**NSD comment**	# EDLVcc was renamed to EDLVAcc and recoded to correspond with categories in questionnaire (ranging from 1-9 rather than 0-8). Please confirm.
	Response from NT	Ok.
EDULVL EDULVLF EDULVLM	**NSD comment**	# Missing codes (77, 88, 99) have been recoded to one-digit level (7, 8, 9) in accordance with data protocol. Please confirm.
	Response from NT	Ok.
Miscellaneous		
IDNO	**NSD comment**	# IDNOs 180 and 192 have more than 50% refusal, Don't know or No Answer (65,8 and 92,8 respectively) in main questionnaire. See p. 47 in Annex. Should these cases be dropped from data set?
	Response from NT	We would prefer to keep IDNO 180. IDNO 192 was an early canceled interview due to language problems. This record can be dropped from the data set.

From round 4, NSD attach round-wise comparison tables to countries that have participated in more than one round of the ESS (figure 3.9). Only variables where distributions have changed significantly are reported, based on a test formula

explained in the figure. This is an extra step in the quality control, and it reduces the work load for the National Coordinators with respect to controlling repetitive variables.

Figure 3.9 Reporting a variable with significant changes in distribution since previous round

B32|Ban political parties that wish overthrow democracy	Round 3	Round 4	absdiff	sum2stdv	alert
1	16.29	12.80	3.49	2.94	1
2	34.38	41.78	7.40	4.04	1
3	21.02	20.91	0.10	3.40	0
4	11.66	10.73	0.93	2.63	0
5	4.23	4.27	0.04	1.68	0
7	0.50	0.30	0.20	0.52	0
8	11.93	9.21	2.72	2.56	1
	100.00	**100.00**			

```
The test is based on the formula for standard error of estimates in
binomial distributions multiplied by 2 stdv: 2*sqrt[(p(1-p)/N)], and
should only be considered as an indicative measure. A variable is only
included in the output if it contains an „alert” set to 1.
<Explanation: 1 = “Agree strongly”  …… 5 = “Disagree strongly”, 7 =
“Refusal”, 8= “Don’t know”>
```

When the draft file has been approved, it is ready for integration in the cross-national file. Since all national files have been processed according to the specifications and standards of the ESS Data Protocol, this step is performed automatically and simultaneously for all countries to be published. In the final step, design weights produced by the ESS Sampling team and population weights, based on official statistics, are included in the data.

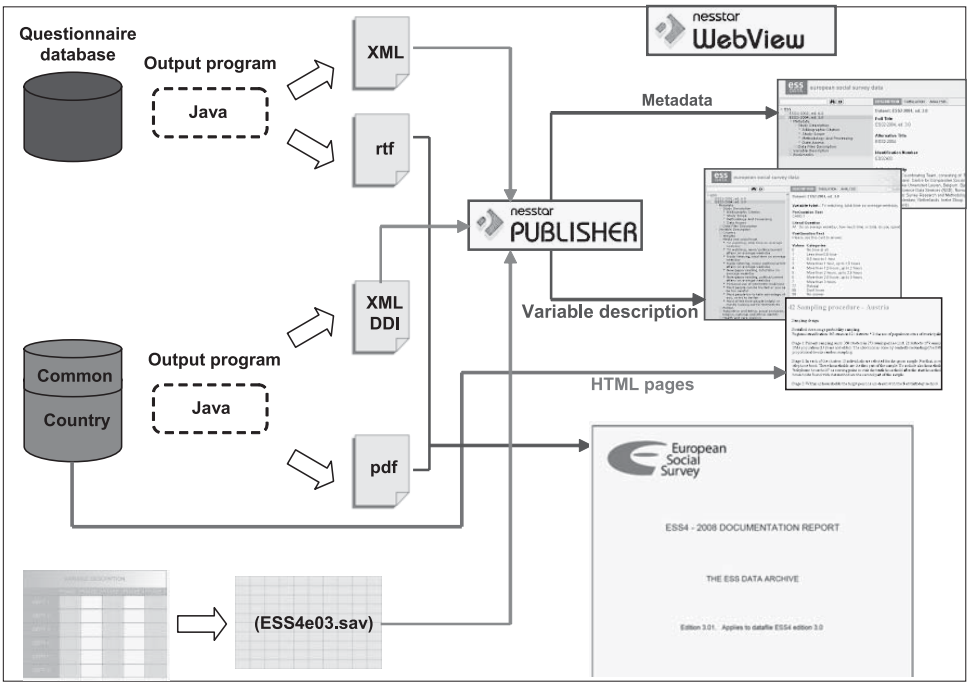
3.3 Processing of metadata

As already mentioned in section 2 the ESS has put large emphasis on providing comprehensive and structured documentation of the survey and the data files. Metadata is about communication and information flow (Ryssevik, 2000). One might argue that the main purpose of metadata is to pass on all relevant information from

one stage of the survey and data life cycle to the next, thus enabling all participants to add their relevant knowledge to this stream of information exchange.

When processing, disseminating and maintaining a complex, long-term survey like the ESS, it is from the data archive's point of view essential also internally to have access to a thorough and comprehensive documentation which is also maintained and updated. The processing of metadata in the ESS is therefore based on generic databases, where documentation, as far as possible, is broken into entities corresponding to the elements of the documentation standard DDI.

Figure 3.10 Metadata work flow in the ESS



As illustrated in figure 3.10, the ESS metadata are stored in a question database and a documentation database. The generic format of the metadata enables us to route the different elements to the relevant parts of the documentation made available to the users, and to easily integrate metadata and data in the ESS Online Analysis tool. And not least; in preparing for and producing cumulative files and time series, standardised metadata from the different cross-sections are indispensable for the archive's work, by providing change history of each variable as well as for elements in the general survey documentation.

The processing of the national metadata collected by the National Technical Summary is of course different from the data processing in nature, but the extent of metadata and the priority given to it, means that the control and processing, registration and maintenance of information in the databases makes up a considerable part of the archive work in this stage of the ESS data life cycle.

3.4 Archiving, preservation, feedback and maintenance

Archiving and preservation

The vast majority of the digital assets of the ESS, including all published material, preliminary data files, processing programmes, raw data etc. have lasting value and are therefore archived in a safe manner. But they must also be preserved for the future. In a long-term perspective digital objects are accessible in their original format for a limited period of time, mainly because of the continual development of computing hardware and software (JISC 2006).

With nearly ten years history, the ESS archive is therefore now focusing on an organisational policy on preservation of the ESS digital assets. Technically, one can distinguish between three classes of strategies for preservation. The first is based on non-digital backups, and are generally not considered as a relevant strategy for objects of digital origin. The second is emulation, which focuses on recreating an original computer environment using current hardware and software. This approach is valuable because of its ability to maintain a closer connection to the authenticity of the digital objects, but it can be complex, time consuming and difficult to achieve. The third class of strategies is migration, where digital objects are converted into current or more widely accessible formats. Digital objects that are migrated run the risk of losing some type of functionality, since newer formats may be incapable of capturing all the functionality of the original format, but is considered as the less complex strategy.

Taking into account that the majority of the digital assets of the ESS consist of digits and text, migration will in general be sufficient, and it is also the most cost-efficient approach. By storing the digital objects in generic formats, as the ESS archive to a large extent already does, the task of preservation of the continuous increasing holdings of data and metadata is considered as manageable.

User feedback and maintenance

Even though data and documentation are handled carefully and with vigilance throughout the data processing, there is always a chance that deviations and errors may prevail in the published dataset or in its documentation. Such errors are often detected through secondary analyses by our more than 30 000 ESS users, and reported back to us. Feedback from users is highly appreciated, as it gives us the opportunity to improve the quality of the data we make publicly available. Thus,

it aids in fulfilling a common goal of the ESS project and the archive per se: To present data of high quality that may be of value to users over time.

The errors detected by users are of various kinds, from errors in variable labels to coding errors, reversed scales, data merging errors, inaccurate or erroneous documentation etc. At NSD, errors in data and metadata are treated according to the nature and scope of the problem and the possibilities for corrections.

Sometimes it is possible to replace erroneous data by correct data provided by the National Teams. One example would be how erroneous merging of post-coded variables into the main data file led to the unexpected finding that no Turkish born respondents in a particular country were found to speak Turkish at home. When notified about this possible error, NSD replicated the analyses and approached the National Team of the country for which the problem had occurred. It turned out that it was possible to retrieve correct data, and erroneous variables in the ESS data file were replaced by the new variables the archive received.

In other instances, correct data are not available, but it may be possible to recode or redress data to improve the quality of the collected data. For example, if a question has been asked with a reversed scale, it is possible to recode the variable in the dataset and document the amendment.

It isn't always possible to replace or redress data. Data could simply not be collected due to routing errors in CAPI programmes. The same occurs when a wrong or deviating instrument has been used in a particular country, and there is no possibility for harmonising these data with the correctly collected ESS data. Feedback regarding irreparable data is still important, as all detected deviations need to be documented and conveyed to the other users of the ESS. The ESS Data Documentation report, as well as the 'Deviations and Fieldwork summary' at the ESS Data website, list all known deviations in the data. Depending on the problem, irreparable data are sometimes removed from the integrated international data file, and made available in a country-specific data file only.

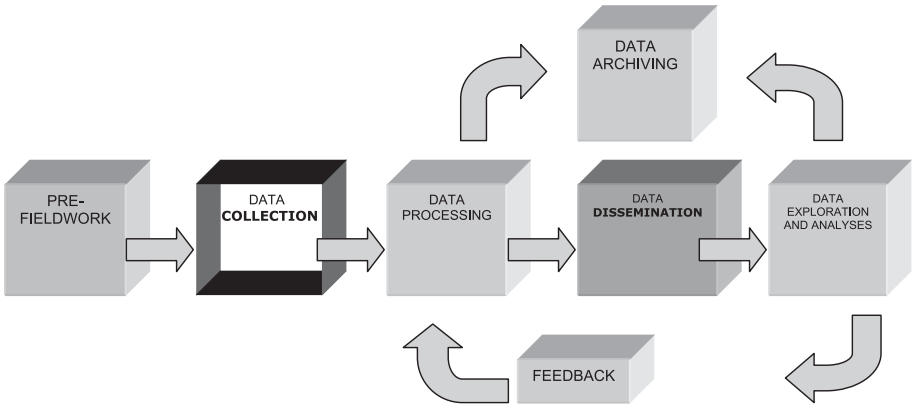
In-depth analyses by users may also reveal weaknesses in the reliability or validity of particular measures. All feedback related to reliability or validity issues are forwarded to the Central Coordinating Team for discussion and decisions on how to improve measurement quality.

As ESS data and metadata grow in amount and complexity over time, a major challenge is to keep the published data available to users as updated as possible. Complete releases with new editions of data and metadata are increasingly time consuming and demanding, and can thus not be performed on a continuous basis. Instead, NSD are looking into more flexible, but also partial releases of corrected variables and documentation. In this way the user need for access to correct data can be maintained, while new releases of the complete data files can be limited to major upgrades.

4. DISSEMINATION[1]

To paraphrase the American author Albert Jay Nock, the business of a contemporary social science data archive is *also* the dissemination of useful knowledge (Nock 1934). The following section will deal with the Dissemination stage of the Data Life Cycle as implemented in the ESS.

Figure 4.1 Data Life Cycle, Dissemination stage



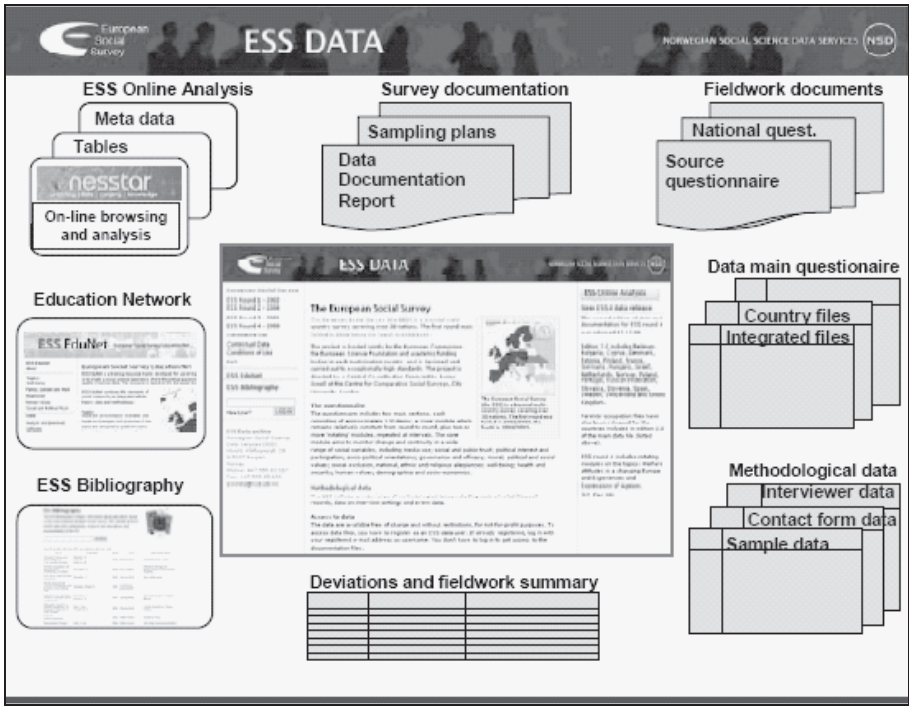
The most common starting point for the data user, even in the recent past, was to have to negotiate a range of institutional distribution practices, such as who is eligible to access the data, when they can do so, and for what purposes. These hurdles were not necessarily substantial – involving perhaps the completion of an electronic questionnaire on the Web or placing an order for a CD-ROM of the data and waiting for it to arrive. But they have generated lots of complaints from data users anxious to get on with their jobs. Some data archives therefore began distributing data over the Internet, providing the data user with an access code. Nevertheless, a "conditions of use" document usually had to be completed and signed in advance of data becoming available. The time involved in gaining access to data by these means naturally varied, but it has always been a bone of contention.

The funding arrangements and the collective nature of the ESS enable the survey to distribute data and documentation to potential users much more swiftly, at no cost, and without the usual legal, institutional and cultural barriers. By using a full range of web-based services for data users, NSD as the ESS Data archive are able to provide not only quicker access to the data, but also access to on-line technology which allows the user to run simple cross-tabulations and to

read the accompanying documentation before having to decide whether or not to download the data. But free and immediate access to the data via the Web through a user-friendly dissemination facility has also resulted in greater equality of access to all comers.

NSD distributes ESS data and documentation by means of two interlinked systems based on Web technology. The overarching system which is visible to the data analyst is the ESS Data website, http://ess.nsd.uib.no. This website serves as the reservoir for the complete set of ESS data, metadata and documentation. Integrated within this website is the on-line analysis and distribution tool Nesstar.

Figure 4.2 Components of the ESS Data website

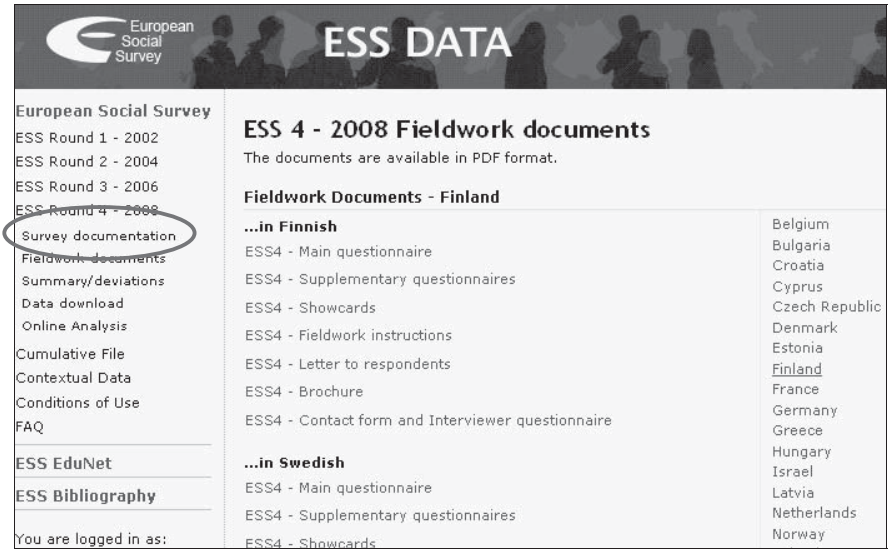


The ESS Data website tries to meet different user requirements and in particular their needs for different levels of documentation. It is divided into sections on data download, fieldwork documents, survey documentation, contextual data, and on-line browsing and download. Starting with the download facility, we briefly describe the role and content of the different sections.

The ***Data download*** section offers access to integrated as well as country-specific files. The integrated files (SPSS and SAS) which are accessible for download hold data from the core and rotating modules of the ESS; the interviewer's questionnaire, which describes the interview setting; the contact forms, which contain contact information on all sample units, including non-response units; and the test questions (used in multi-trait multi-method (MTMM) analysis of reliability and validity).

The integrated main data file, containing the core and rotating modules, is also available as separate country files. The website also provides access to data files with country-specific variables, such as variables omitted because they were not collected in an equivalent way, or extra variables that were included only in a specific country. For instance, the country-specific data file from Germany in ESS round 4, contains all (20) variables on education level that were used to bridge from the national measurement to the standardised ESS variable (EDULVL). On country level, users can also find data files recording parents' occupation in local languages and for some countries also as ISCO codes.

Figure 4.3 Access to fieldwork documents, the ESS Data website



The ***Fieldwork documents*** section contains questionnaires, show cards, contact forms and fieldwork instructions in all languages (see figure 4.3). So users can simply select a country and bring up country-specific versions of these source documents. Other documents used during fieldwork, like advance letters and

brochures are also made available. This allows data users to access the translated questionnaires and assess for themselves if the translation was functionally equivalent, as compared with the source questionnaire. National documents are of course an indispensable source in the disclosure of flaws that have to be corrected in future rounds of the ESS. All deviations in the data are thus documented on-line and in detail in a Summary/deviations section. By making all these documents freely available online, not only the data user, but also the wider survey community, may evaluate the survey's quality and thus be able to replicate (or avoid) certain aspects in future projects.

The ***Survey documentation*** section contains documents aimed primarily to assist users in the analysis of data. It includes guidelines on how to apply the weights included in the data files, documentation of the national sampling plans and design weights, and reports on question reliability and validity. Survey documentation is found in the *ESS Documentation Report*, which in turn contains three main sections. The first section includes the overall study description, that is information about the study itself, its key people and institutions, how to access its data, and a summary description both of the data file and the legal aspects of data use. The second section contains country-specific details, such as fieldwork agencies, funders, sampling and fieldwork procedures, response rates, and so on. The third section is organised as separate appendices, containing country-by-country population statistics, documentation of classifications and standards used in the survey, plus a list of variables and questions in the main and supplementary questionnaire and variable lists sorted by question number and variable name. From ESS round 1 to ESS round 4, the number of appendices to the Documentation Report has increased from 4 to 6, reflecting the dynamic nature of a multi-national survey such as the ESS and the increased demand for documentation as the survey evolves.

The ***Contextual data*** section currently contains a link to The Macro Data Guide, which originates from a report prepared for the European Social Survey, examining the availability and comparability of extant sources of contextual statistics of particular interest to users of ESS Data.

Finally, the ***On-line browsing and download*** section provides access to Nesstar (see figure 4.4).

As may be seen from Figure 4.4, the left window of the screen is a hierarchy of folders containing various kinds of information and documentation. The right window then shows the selected item in the hierarchy. The two parts of the screen also provide details of the metadata, the keywords, the topics and the abstract.

The hierarchy of the browser tree is organised in two main folders with sub-folders each containing several elements of documentation. The two main folders contain metadata on the one hand and variable descriptions on the other. The system is highly dependent upon structured documentation. In fact, the

Figure 4.4 Nesstar On-line browsing and download, Study Scope ESS4

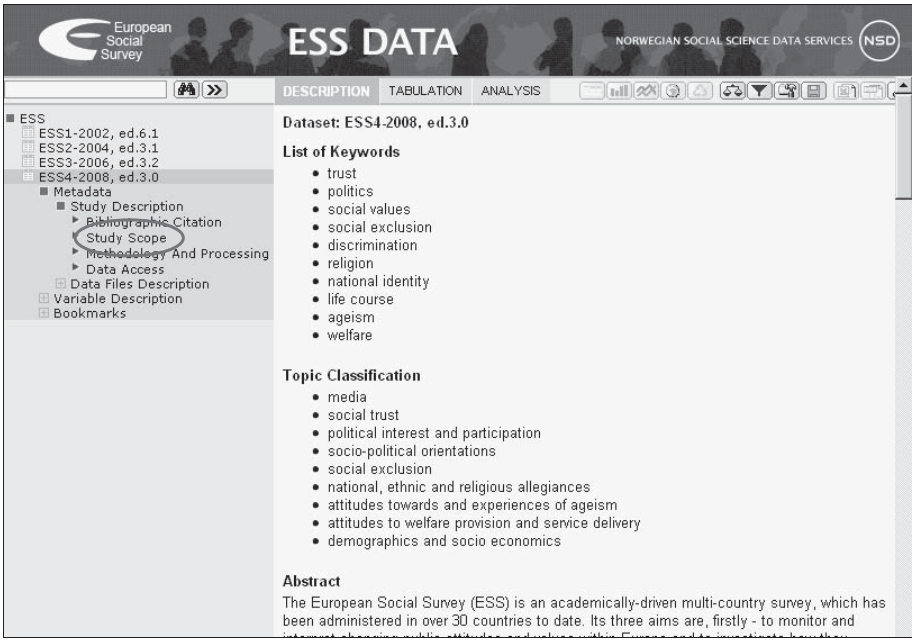
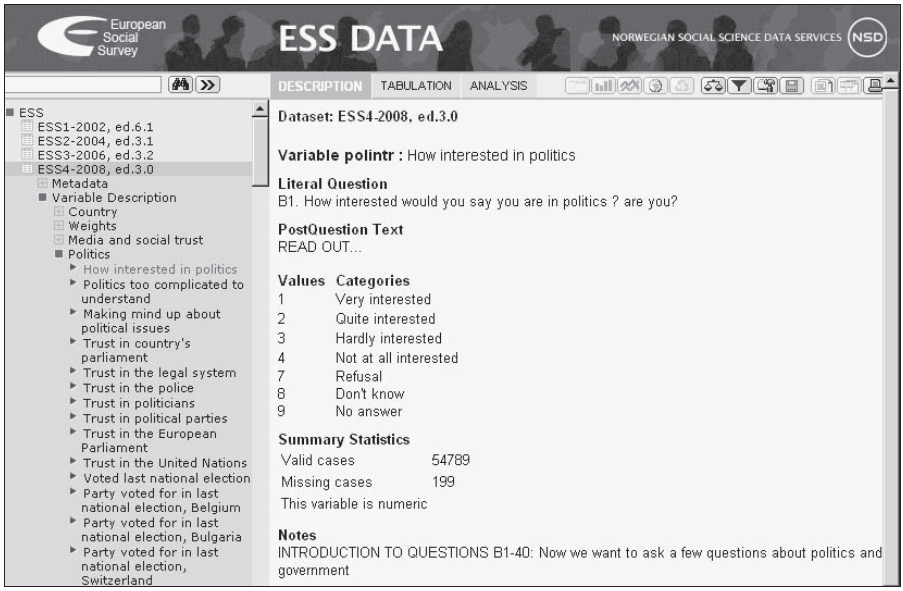


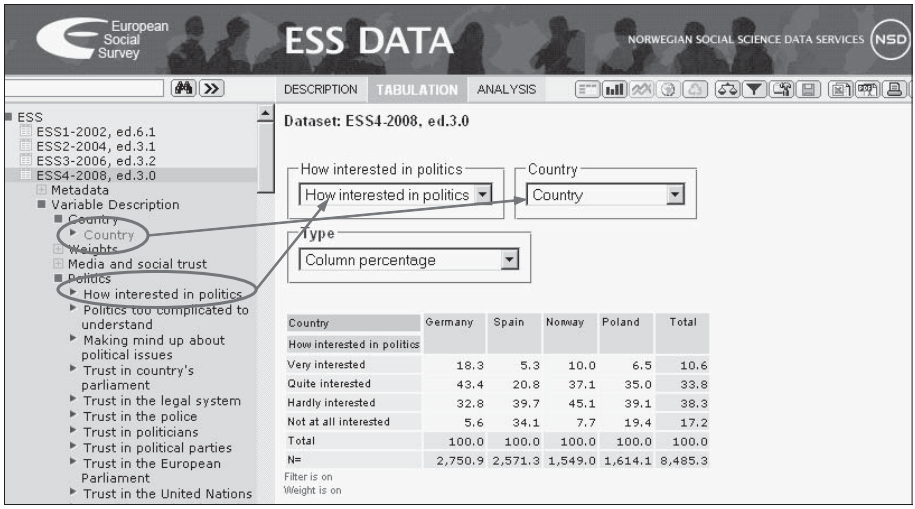
Figure 4.5 Nesstar Variable Description, interest in politics



Nesstar system is built on the standardised tag library Document Type Definition (DTD), developed by the Data Documentation Initiative (DDI). The metadata folder expands to a series of sub-folders, containing both country-specific and survey-level information on different aspects of the survey and its procedures. The variable description folder contains sub-folders, representing the different thematic sections of the questionnaire. Thus, having selected a particular variable, the documentation will be displayed in the window to the right (see Figure 4.5).

As can be seen in figure 4.5, the listing on the right includes the full question asked, together with all answer categories, plus any interviewer instructions for the question. It also displays the variable name, its position in the questionnaire (first question in section B), summary statistics about the variable (valid/missing cases), and any notes or warning related to that question. An option also exists to obtain a tabular view of the distribution of the variable.

Figure 4.6 Nesstar crosstabulation of a weighted subset (ESS4)

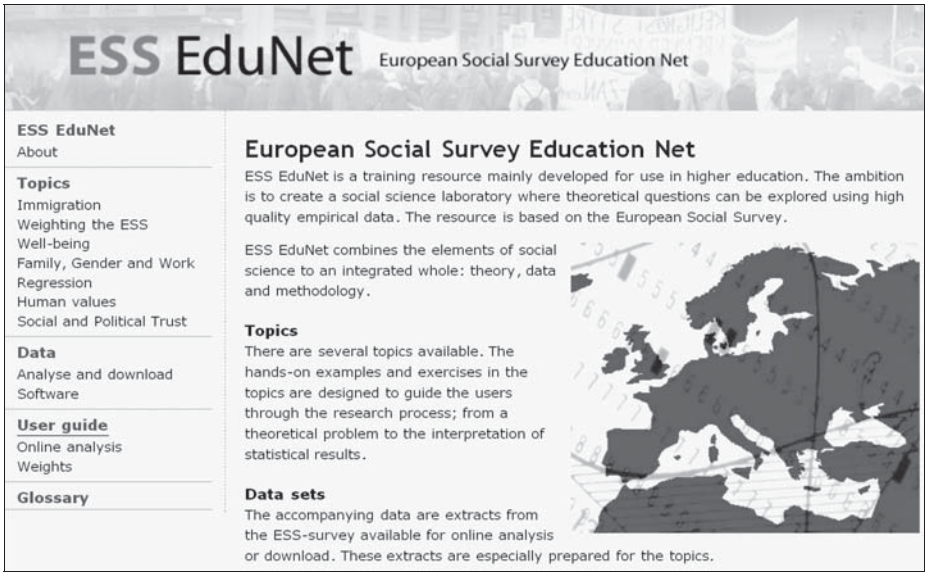


Thus, the Nesstar system serves in effect as an electronic codebook. But it not only makes it possible to display the distribution of the variables, but also to select them and display them in cross-tabulations (see figure 4.6). Descriptive statistics like mean, median, standard deviation, quartiles etc. are also available options, along with correlation and simple regression. The tables can also be graphically represented for example in bar charts. In datasets containing geographical identifiers, maps can be added to the resource, enabling the data to be displayed spatially. An additional feature of Nesstar, is the ability to bookmark any output

(table or analysis) produced. These bookmarks can then be kept for later use or shared with others by providing them the link to the output.

Another useful feature of the system, is its ability to create subsets of the data, based on the values of a particular variable. Thus, a subset of certain countries or individuals can be created based on the similarity of their characteristics (in full-time work etc). It is also possible to carry out simple regressions, and the system offers a weighting option and the possibility of displaying tabular information graphically without having first to download the data. Whether for analysis, or simply as a means of tailoring a specific subset of the ESS, the data may, however, also be downloaded in a number of different formats, including SPSS, SAS and Stata.

Figure 4.7 ESS EduNet

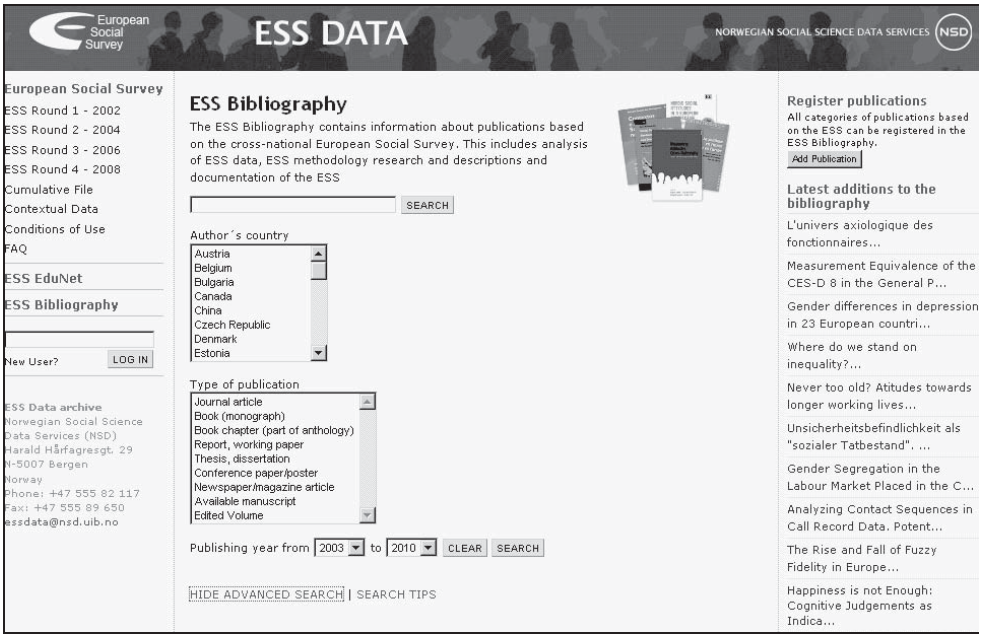


ESS EduNet also provides an important dissemination tool for the ESS. The aim of the ESS EduNet is to provide a pedagogical resource that makes it simpler for the coming generations of researchers and higher level students in the social sciences to utilise the ESS data. It also provides lecturers with ready-made ESS topics to use in their teaching. The resource currently contains seven modules (see figure 4.7), each based on the data that emerged from one of the substantive modules within the three first rounds of the ESS. Each module refers to a limited subset of ESS data, carefully selected to focus on a manageable set of dependent and independent variables. In each case, the background theory behind the module

is summarised in the package. The exercises are designed to be solved in part online (using Nesstar technology) and in part using downloaded data, processed locally in more sophisticated statistical packages. Scholars in the substantive fields have produced the modules in cooperation with NSD.

Finally, the ESS Bibliography is an important resource in documenting and disseminating the output of ESS based research, enhancing the visibility of the ESS. The ESS Bibliography is an on-line dynamic service for registration and retrieval of bibliographical information about the usage of the ESS. It provides flexible access and retrieval. It also offers efficient search for all publications that are available on-line. The system is freely available, providing researchers, governments, organisations, decision makers and others with a unique list of social reporting on Europe. The bibliography database can also be used to produce publication lists by country, publication type, year of publication etc., documenting the outcome of the ESS on a detailed level (see figure 4.8).

Figure 4.8 Search options in the ESS Bibliography



5. CHALLENGES FOR THE FUTURE

The ESS has been a huge success. Four biannual rounds with substantial as well as methodological data from an average of 25 countries and 50.000 respondents

have been disseminated to an impressive number of end users. 32.000 individuals are registered in NSD's ESS user database. More than 20.000 of these have downloaded one or more data files from the ESS Data website, and an unknown number of users have accessed data from national ESS websites. Inherent in this success is an increasing amount and complexity of the ESS, and this is a challenge for all aspects of the Data Life Cycle. Not only does the number of participating countries in the ESS increase over time, so does the scope of the specifications and instructions to be conveyed to the National Teams. Thus, good communication between the archive, the National Teams and the Central Coordinating Team is vital for the future success of the ESS.

Robustness and flexibility in underlying systems and databases will be decisive in order to maintain and improve the services provided to the user community and to the ESS partners. As indicated in section 3, keeping track of change history in data and metadata will, in our view, have to be based on robustly designed databases, where questions and other metadata units are treated as generic and relational objects. Question wordings, instructions etc. must for example be linked to the questionnaires, as well as to variables and all their attributes, and thereby indirectly to data. Improving the databases and the underlying systems is also an important step with respect to efficiency and costs. Just as important for efficiency, is to keep focus on the procedures of data processing and to use experiences gained through four rounds of ESS to improve these.

Considering the development over the last decade, where the Internet has gained a key role in the dissemination of practically everything, including survey data, we see no alternative means of distribution that can better cater for the ESS key objective of being easily and freely available for all. However, the visibility in the increasingly diverse landscape of projects and datasets available on the Internet will become a challenge. Even if more than 30.000 individuals are registered in the user database and the ESS is used extensively in higher education in Europe, it will be important that users easily can find the ESS through search engines and other Internet tools. One important way to achieve this, is the establishment of national ESS websites that are closely linked to the authoritative and complete official ESS Data website. These local websites might also contribute to break down language barriers for users not so accustomed to English.

With respect to the dissemination of the ESS, whether it is data, metadata, specifications, or research output, the large amount and diversity as such could also create new barriers for the users. The ESS and NSD will therefore give high priority to improve the structure and design of the main dissemination channels, the websites. In our view, the best way to maintain and increase user friendliness in the future, will be to offer ESS data and metadata in an interactive environment, where users can search for, select and download exactly the data and documentation they

need, be it single cross-sectional data or trend data from all rounds of the ESS. To provide such services in a satisfactory manner, we need close contact with users, and facilitate for user feedback by way of data user forums, usability tests and extensive user surveys.

NOTES

1 This section is based on Kolsrud, Skjåk, Henrichsen (2007).

REFERENCES

Dale, Angela and Ugo Trivellato. 2002. *Access to Microdata for Scientific Purposes.* Background paper to the 19th CEIES Seminar. Lisbon 26–27 September 2002.

JISC 2006. Digital Preservation briefing paper, http://www.jisc.ac.uk/publications/documents/pub_digipreservationbp.aspx

Kolsrud, Kirstine. 2007. *The Work and Challenges of the Data Archive.* Presentation at the Second ESRA Conferenc. Prague June 25–29 2007.

Kolsrud, Kirstine. 2009. *Access to Survey Data on the Internet.* Presentation at the Third ESRA Conference. Warsaw June 29–July 3 2009.

Kolsrud, Kirstine and Knut Kalgraff Skjåk. 2004. *Harmonising Background Variables in International Surveys.* Paper to the RC33 Sixth International Conference on Social Science Methodolog. Amsterdam 16–20 August 2004.

Kolsrud, Kirstine, Knut Kalgraff Skjåk, and Bjørn Henrichsen. 2007. 'Free and immediate access to data.' In: *Measuring Attitudes Cross-nationally.* Roger Jowell, Caroline Roberts, Fitzgerald Rory, Eva Gillian (eds.). London: Sage.

Miller, Warren E. 1976. 'The Less Obvious Functions of Archiving Survey Research Data.' *American Behavioral Scientist*. March/April 1976, Vol 19 nr 4. Sage Publications.

Nock, Albert J. 1934. 'The Value of Useless Knowledge,' *Atlantic Monthly,* May 1934.

Mohler, Peter Ph. and Rolf Uher. 2003. 'Documenting Comparative Surveys for Secondary Analysis.' In: *Cross-Cultural Survey Methods*. Janet A. Harkness, Fons J.R. Van de Vijver, and Peter Ph. Mohler (eds.)*,* New York: Wiley.

Rokkan, Stein. 1976. 'Data Services in Western Europe.' *American Behavioral Scientist,* March/April 1976 Vol 19 nr 4. Sage Publications.

Rokkan, Stein and Bjørn Henrichsen. 1976. 'Building Infrastructures For the Social Sciences: The Norwegian Social Science Data Service.' In: *Research in Norway 1976.* Mary Bjærum, Kari Kristiansen, and Arne Sundland (eds.), Oslo 1976.

Ryssevik, Jostein. 2000. Bazar Style Metadata. In *The Age of the Web – An Open Source Approach to Metadata Development.* Working paper No. 4. UN/ECE Work Session on Statistical Metadata Washington D.C.; United States, 28–30 November 2000.

Skjåk, Knut Kalgraff. 2007. *Clean Data or Cleaned Data? Data Editing Procedures and Experiences of the ESS Data Archive.* Presentation at the Second ESRA Conference, Prague June 25–29 2007.

UKDA homepage http://www.data-archive.ac.uk/about/faq.asp#arc

Kirstine Kolsrud, is Senior Advisor at the Norwegian Social Science Data Services Ltd. (NSD), where she has worked since 1993. She gained her graduate degree in Economics in 1995, after which she carried out econometric analysis and desk research at the Department of Economics, University of California, San Diego (1996). At NSD, she has worked with databases, individual level data from official registers, privacy issues and developed performance indicators on council services such as primary education, social services, and local health services. She has also taken part in the planning and work of cross national surveys at NSD, such the Health Behaviour among School Children (HBSC), the International Social Survey Programme (ISSP) and the parallel surveys to the Eurobarometers in Norway. She has been a member of the ISSP Secretariat at NSD from 2003 to 2009. Kirstine Kolsrud has been working on the European Social Survey since 2001. She is the project coordinator for the ESS archive work at NSD and is a member of the Central Coordinating Team.
E-mail: essdata@nsd.uil.no

Hege Midtsæter, has worked as a Specialist Consultant at NSD since 2010. She gained her masters degree in Economics and Business Administration in 2009. At NSD she is working with the cross-national surveys the European Social Survey (ESS) and the International Social Survey Programme (ISSP).
E-mail: essdata@nsd.uil.no

Hilde Orten, is Advisor at NSD, where she has worked since 2002. She gained her BA in Philosophy in 1986 and her graduate degree in Sociology in 2007. At NSD, she has worked with cross-national surveys. Hilde Orten has worked on the European Social Survey since 2002. She has also taken part in the work of other cross national surveys at NSD, such as Health Behaviour among School Children (HBSC), the International Social Survey Programme (ISSP), and the parallel surveys to the Eurobarometers in Norway. She has also been involved in the planning of a new infrastructure on data harmonisation for the Council of European Social Science Data Archives (CESSDA).
E-mail: essdata@nsd.uil.no

Knut Kalgraff Skjåk, is Assistant Director at NSD and head of the External Projects Division. He gained his graduate degree in Social Geography in 1983. He has been director of the Norwegian part of the International Social Survey Programme (ISSP) since 1989, and member of the ISSP Secretariat at NSD from 2003 to 2009. He has directed a number of national surveys on social values and attitudes, and implemented and supervised various methodological experiments in survey research in Norway. He has been responsible for the international data archive of Health Behaviour among School Children (HBSC) at NSD and has supervised parallel surveys to the Eurobarometers in Norway. He has been working on the European Social Survey since 2001, and is a member of the ESS Central co-ordinating team.
E-mail: essdata@nsd.uil.no

Ole-Petter Øvrebø, is a Specialist Consultant at NSD, where he has worked since 2007. He gained his graduate degree in Comparative Politics in 2003. Prior to joining the European Social Survey team at NSD, he enjoyed a short spell as Research Assistant at the Department of Comparative Politics at the University of Bergen, as well as a couple of years in the field, representing Norwegian authorities in the Balkans as a political analyst.
E-mail: essdata@nsd.uil.no